DOCUMENT RESUME

ED 218 300 TM 820 327

AUTHOR Holland, Paul W.; Rubin, Donald B.
TITLE Causal Inference in Prospective and Retrospective Studies.
SPONS AGENCY Educational Testing Service, Princeton, N.J.; National Inst. of Education (ED), Washington, DC.
PUB DATE Aug 80
GRANT NIE-G-78-0157
NOTE 47p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Mathematical Models; Research Design; Research Methodology; Statistical Analysis; *Statistical Studies
IDENTIFIERS *Causal Inferences

ABSTRACT

Emphasizing the measurement of causal effects to arrive at a better understanding of the causal mechanisms involved in statistical theory, a mathematical model for causal inferences in prospective studies is developed and then applied to retrospective case-control studies. Before developing the model, causal agents are delineated, and causal effects are distinguished from "gains over time". The formal model is presented considering indirect measurement of causal effects, homogeneous populations, intermediate-level causal effects, the selection variable, randomization and the role of covariates. In the retrospective case-control studies, retrospective and prospective probabilities and matching are discussed. A loglinear model for a case-control study problem is presented. (Author/CM)

Causal Inference in Prospective and Retrospective Studies

by

Paul W. Holland

and

Donald B. Rubin

August, 1980

This work was partially supported by the Program Statistics Research Project at Educational Testing Service. Paul Holland was partially supported by grant NIE-G-78-0157 from the National Institutes for Education. Donald Rubin's work was facilitated by a John Simon Guggenheim Fellowship.

Preface

It is an honor to present this discussion paper at the Jerome Cornfield Memorial Session of The American Statistical Association. The topic of our paper seems especially appropriate for this session since many important contributions to the study of health effects from prospective and retrospective studies were made by Jerome Cornfield, e.g. Cornfield (1951, 1956).

1. Introduction

Philosophical discussions of causality can be far ranging and touch upon an enormous variety of subjects. The reason is the emphasis, in the philosophy of science, on the understanding of causal mechanisms. Statistical discussions of causality are substantially more limited in scope because the contributions of statistics are to the measurement of the size of causal effects and not to the understanding of causal mechanisms. This distinction is sometimes expressed as "statistics can establish correlation but not causation." We feel our emphasis on measurement versus understanding is more appropriate because it focuses on the things that statistical theory can contribute to discussions of causality rather than on what it can not. It is perfectly possible to measure a causal effect accurately without any understanding, whatsoever, of the causal mechanisms involved. The measurement of causal effects without understanding the causal mechanism is, of course, a commonplace experience of everyday life, i.e., people are quite capable of using automobiles, ovens, calculators and typewriters safely and effectively without detailed or, in some cases, any knowledge of how these devices work. On the other hand, precise measurement of causal effects often leads to a better understanding of the causal mechanisms involved.

In this paper we develop a mathematical model for causal inferences in prospective studies that is based on the work of Rubin (1978) and we then apply it to causal inference in retrospective case-control studies. Before developing this model, we shall briefly delineate what we consider to be properly called causal agents in a statistical discussion, and shall also sharply distinguish causal effects from "gains over time" -- two ideas that our experience suggests are often confused with each other.

Causal Agents

Consider the following two statements:

(a) "That person didn't do well on the exam because she did not study first."

(b) "That person didn't do well on the exam because she is a woman."

In statement (a), the implied causal agent is the amount of studying; that is, had the person studied harder, she could have done better on the test. In other words, there was a point in time when a choice was made either to study or not to study and the comparison between the subsequent scores on the exam is the causal effect of studying versus not studying.

Statement (b) is statistically very different from statement (a) in that there is no choice of levels of a causal agent possible, i.e., the person cannot choose or be assigned to be a male or female. Consequently, there is no logical comparison possible for a single individual between their score on the test as a female and their score on the test as a male. The use of cause in statement (b) reflects only the correlation between attributes of individuals. In medical studies the term "risk factor" is sometimes used broadly to encompass both causal agents like smoking which can be altered and individual attributes like age and sex which can not. The identification of gender or other individual attributes such as race as a causal agent in such questions is statistically meaningless. It may also distract scientists from the study

of causal agents that can have beneficial effects, e.g., finding programs of study that are particularly effective for women and others that are particularly effective for men.

It is common usage to say that the levels of causal agent are treatments, especially when their assignment is under an experimenter's control.

Our definition of causal agent is much stricter than some definitions commonly used by economists, e.g. Granger causality (Granger, 1969). We believe that to lable any successful predictor a causal agent not only misuses the language and thus is deceptive, but also may lead researchers away from study of the relevant scientific questions of the effect of manipulations that are possible.

Gains versus causal effects

In order to distinguish between gains and causal effects, consider a student who was coached for the Scholastic Aptitude Test (SAT) between the first administration of the SAT, and the second. Let the two scores be SAT_1 and SAT_{2C} -- the subscript C indicating the coaching that took place between administrations of the SAT. The causal effect of coaching is not $SAT_{2C} - SAT_1$. This difference is the gain (or loss) over time in SAT scores. The causal effect of coaching is the difference between the SAT score at administration 2 given exposure to coaching, SAT_{2C}, and the SAT score at administration 2 given no exposure to coaching, say SAT_{2N}. The pre-coaching SAT score, SAT_1, may be useful in estimating the causal effect on coaching, but $SAT_{2C} - SAT_1$ is not equal to the causal effect of coaching unless we assume that $SAT_1 = SAT_{2N}$, i.e., a "no change without coaching" assumption. The tenability of such "no change" assumptions in general depends a great deal on the particular substantive problem under study. It is probably false in this SAT example.

2. Causal inference in prospective studies

The logic of measuring the size of causal effects is clearest in prospective studies, and so we shall begin with that case. The essential elements of a prospective study are the following:

(a) a population of units, Q

(b) a set of well-defined levels of a causal agent (or treatments) to which each unit Q could be exposed. (For notational simplicity, we consider only two treatments denoted by $\ell=1$ or $\ell=2$)

(c) a response Y which can be recorded for each unit after exposure to a treatment.

In a prospective study, a sample of units from Q is obtained and the units are assigned to treatments. The treatments are then applied, and later the response of each unit in the study is recorded. The intuitive notion of causal effect that we wish to describe with our model is the difference between the response measured on a unit that is exposed to treatment 1 and the response that would have been measured on the same unit had it been exposed to treatment 2. Thus, our notion of the causal effect of a treatment will always be relative to another treatment, and is defined for each unit in Q.

This meaning of causal effect is not foreign to statistical thinking and is evident in the writings of R.A. Fisher (1925), Kempthorne (1952), Cochran (1965), and Cox (1958), for example. Although this notion of a causal effect can be defined for each unit in Q, in general, we are not able to directly measure a causal effect for a single unit because having given treatment 1, we cannot return in time to give treatment 2 instead. This is the fundamental problem of causal inference, and our

formal model will show how its solution is related to the use of randomization and of covariates.

Before turning to the formal model we need to define the nature of the response Y. For our discussion we will assume that Y is dichotomous, taking on only the values k=0 or k=1. The extension to Y taking values in an arbitrary finite set is straightforward. We have chosen to restrict Y to be discrete in order to emphasize the fundamental ideas behind the measurement of causal effects without being distracted by the special mathematical baggage that is automatically associated with continuous variables -- i.e., additivity, etc.

The formal model and definition of unit-level causal effect

In our model, instead of a single dependent variable Y, we have a dependent variable, Y_ℓ for each of the treatments to which the unit could have been exposed. Thus, if the unit is exposed to treatment 1, then we will record the value of Y_1 for that unit. If that same unit had been exposed to treatment 2 instead of treatment 1, then we will record the value of Y_2 for that unit and not the value of Y_1. More formally, for two treatments, with each unit in Q we associate the following partially observable vector of information:

$$(Y_1, Y_2) \qquad (1)$$

where

Y_ℓ = response made by the unit if it is exposed to treatment ℓ.

The novel feature of this model is the introduction of several versions

of the response variable, Y. There is a version of Y for each level of the causal agent because our definition of causal effect compares Y_1 (the response made if exposed to level 1) to Y_2 (the response made if exposed to level 2). The fact that each unit has a value for both Y_1 and Y_2 is very important because it allows us to <u>define</u> causal effects at the level of individual units. If $Y_1 = 1$ and $Y_2 = 0$ for a particular unit, then the causal effect of treatment 1 relative to 2 for that unit is to change the response for that unit from 0 to 1. Rubin (1980) refers to the assumption that the vector (1) fully represents the possible values of Y under all treatment assignments as the "stable unit-treatment value" assumption.

A question that immediately arises is whether or not it is <u>ever possible</u> to expose a unit to more than one treatment and thereby directly observe more than one component of the vector in (1). One can argue that this is never possible in principle, because once a unit has been exposed to a treatment, the unit is different from what it was before. However, the reasonableness of this extreme position depends on the nature of the treatments and on the units under study. We will not pursue this issue further here but will simply make the "worst case" assumption that a unit can be exposed to at most one treatment condition. For our application to retrospective studies this assumption is adequate since in these studies units are only exposed to one level of treatment.

In going from the partially observable vector in (1) to the observable data we must introduce the variable S where $S = \ell$ if

the unit is exposed to treatment ℓ; S is the "selection" or "assignment" variable that indicates to which treatment the individual is exposed.

The observable data from a unit in Q is the vector

$$(Y_S, S) \tag{2}$$

The notation Y_S is used because it indicates that we can observe only the response of a unit to the treatment to which it is exposed, i.e.,

$$Y_S = Y_\ell \text{ if } S = \ell, \quad \ell = 1,2 \tag{3}$$

The quantity Y_S is the observed value of the response and is therefore what is usually called the "dependent variable" in statistical discussions. We never get to observe Y_ℓ if $S \neq \ell$. Since we can observe only the value of Y_1 or Y_2 but not both, it is a consequence of the model that causal effects for individual units are not directly measurable. Indirect measurement of the causal effects is sometimes possible, and our purpose here is to analyze this possibility for both prospective and retrospective studies.

In summary, our idealized model for a prospective causal study can be viewed as based on the following sequence of steps.

(A) Determination of the population Q under study.

(B) Determination of the treatments under study.

(C) Determination of the response variable Y to be observed.

(D) Consequent definition of the vector (Y_1, Y_2) for every unit in Q.

(E) Determination of the assignment variable S for every unit in the study.

(F) Consequent definition of the vector (Y_S, S) for every unit in the study.

(G) Observation of (Y_S, S) for each unit in the study.

Indirect measurement of causal effects

Although our definition of causal effect at the unit-level corresponds to most everyday uses of the term "cause" (e.g.; I didn't do well on the exam because I didn't study), scientific studies often must be content with measuring a weaker notion of causal effect. In the population Q, suppose there are $n_k(\ell)$ units for which $Y_\ell = k$ $\ell=1,2$; $k=0,1$. That is, $n_0(1)$ is the number of units for which $Y_1=0$. If n_+ denotes the total number of units in Q, then the vector

$$q(\ell) = (q_0(\ell), q_1(\ell))' = (n_0(\ell)/n_+, n_1(\ell)/n_+) \qquad (4)$$

gives the distribution of responses under treatment ℓ for the entire population Q. A weaker definition of causal effect of treatment 1 relative to 2 is based on the comparison of the two response distributions $q(1)$ and $q(2)$. If, for example, $q_1(1) > q_1(2)$, then the population-level causal effect of 1 relative to 2 is to increase the proportion of units in Q for which Y=1. We shall call $q(\ell)$ the causal parameters of the study. In terms of the distribution of Y_ℓ over Q $q_k(\ell)$ may be expressed as

$$q_k(\ell) = P(Y_\ell = k). \qquad (4a)$$

Consider a simple randomized experiment. A random sample of units from Q are exposed to treatment 1 and the values of Y_1 are obtained for them. This gives us an estimate of q(1) which has an accuracy that depends on the size of the random sample. A second random sample of units from Q is exposed to treatment 2 and the values of Y_2 are obtained for them. This yields an estimate of q(2). A comparison of these two estimated causal parameters is a form of causal inference because it allows us to say that treatment 1 causes a change in the entire distributi of responses for the units in Q relative to the distribution of responses under treatment 2 by a given estimated amount.

Homogeneous populations

A population-level causal inference is weaker than a unit-level causal inference because it does not allow us to say how treatments change the response of any single unit in Q except in one very special and important circumstance which we now discuss. If Q is such that Y_1 takes on a single value for all units and Y_2 also takes on a single value (that is possibly different from that of Y_1) then Q will be said to have homogeneous responses for treatments 1 and 2. We shall refer to such a Q as a "homogeneous population". When Q is a homogeneous population, then the population-level causal inference is equivalent to unit-level causal inferences for all the units in Q. For example, if

$$q(1) = (q_0(1), q_1(1)) = (0,1)$$

and if

$$q(2) = (q_0(2), q_1(2)) = (1,0)$$

then treatment 1 changes the responses of every unit in Q from Y=0 under treatment 2 to Y=1.

Earlier we distinguished between individual attributes and causal agents. Attributes can be used to partition Q into subpopulations. Finding homogeneous subpopulations plays an essential role in much of scientific research. In the physical sciences the search for "identical initial conditions" is really the search for collections of units (i.e., populations) with homogeneous responses. An "ideal covariate" is an attribute (or set of attributes) which may be observed for each unit in Q prior to the onset of the treatments and which defines subpopulations of Q, each of which has homogeneous responses to the relevant treatment conditions. In practice, of course, we must often settle for less-than ideal covariates which only define subpopulations that are relatively homogeneous.

Intermediate-level causal effects

There is an intermediate level between unit- and population-level causal inferences. Consider all of the units in Q which respond with the value k under treatment 2. We may ask, in what way does treatment 1 change the responses of these units? That is, what is the distribution of values of Y_1 for the units in Q with $Y_2=k$? The answer to this question is a more detailed causal inference than a population-level causal inference, and yet it aggregates units in Q so that it is less detailed than a unit-level causal inference. This intermediate-level causal

inference leads naturally to the notion of a causal-effect table for Q. Let

$n_{k,k'}$ = number of units in Q for which $Y_1=k$ and $Y_2=k'$.

Since n_+ is the total number of units in Q,

$$q_{k,k'} = n_{k,k'}/n_+ \qquad (5)$$

is the proportion of units in Q for which $Y_1=k$ and $Y_2=k'$. In terms of the joint distribution of Y_1 and Y_2 over Q $q_{k,k'}$ may be expressed as

$$q_{k,k'} = P(Y_1=k, Y_2=k'). \qquad (5a)$$

Let $\underset{\sim}{q}$ be the 2x2 matrix with entries $q_{k,k'}$. Then the row totals of $\underset{\sim}{q}$ yield the distribution of responses under treatment 1, i.e., $\underset{\sim}{q}(1)$, and the column totals of $\underset{\sim}{q}$ yield the distribution of responses under treatment 2, i.e., $\underset{\sim}{q}(2)$. We call $\underset{\sim}{q}$ the causal-effect table for treatments 1 and 2 in Q. Table 1 is a causal-effect table.

Table 1 about here

As discussed earlier it is often possible to estimate the marginal distributions $\underset{\sim}{q}(\ell)$, $\ell=1,2$ using randomization. However, it is generally not possible to estimate the joint distribution $\underset{\sim}{q}$. This problem arises because of our fundamental assumption that Y_1 and Y_2 can never be simultaneously observed on any unit. The one situation in which $\underset{\sim}{q}$ can be estimated arises when Q is a population with homogeneous responses. The causal-effect table for a homogeneous population is illustrated in Table 2 and we see there that $\underset{\sim}{q}$ is uniquely determined by the marginal distributions $\underset{\sim}{q}(1)$ and $\underset{\sim}{q}(2)$.

Table 2 about here

Table 1

Causal-effect Table for Treatments 1 and 2 in a Population

Y_1-values	Y_2-values: 0	1	Total
0	q_{00}	q_{01}	$q_0(1)$
1	q_{10}	q_{11}	$q_1(1)$
Total	$q_0(2)$	$q_1(2)$	1

Table 2

Causal-effect table for a population that has homogeneous responses under treatments 1 and 2, $Y_1 = 0$, $Y_2 = 1$.

		Y_2 = 0	Y_2 = 1	Total
Y_1	0	0	1	1
	1	0	0	0
Total		0	1	1

When Q is not homogeneous, it may be possible to decompose it into homogeneous subpopulations, and compute the causal-effect table for each of these subpopulations. It is then possible to accumulate these subpopulation causal-effect tables to obtain the overall causal-effect table for Q. If it is not possible to find homogeneous subpopulations of Q then it is not possible to form the causal-effect table for Q from its margins because the entries are not determined by q(1) and q(2).

Since we rarely encounter perfectly homogeneous populations in practice, we may raise the question of how constrained is q if we only know (or can estimate) the causal parameters q(1) and q(2). The kinds of constraints that exist are easily conveyed by a few examples; these are given in Table 3. The margins of these causal effect tables are considered to be known and fixed, and the range of possible values for the cell entries are given in parentheses. It is evident that if one of the cells in each margin is near one, q is highly constrained. When none of the proportions in q(1) and q(2) is large, q is less constrained. The general rule for calculating the ranges of values for these tables is given by:

$$\max(0, q_k(1) + q_{k'}(2) - 1) \leq q_{k,k'} \leq \min(q_k(1), q_{k'}(2)) \quad (6)$$

The selection variable and the role of randomization

The causal-effect table gives the joint distribution of Y_1 and Y_2 over Q. The data in a causal effect study consists of values of (Y_S, S) for each unit in the study. The joint distribution of (Y_S, S) over Q does

Table 3

Y_1 \ Y_2	0	1	
0	(0,10)	(80,80)	90
1	(0,10)	(0,10)	10
	10	90	

Y_1 \ Y_2	0	1	
0	(40,50)	(40,50)	90
1	(0,10)	(0,10)	10
	50	50	

Y_1 \ Y_2	0	1	
0	(0,50)	(0,50)	50
1	(0,50)	(0,50)	50
	50	50	

not determine the joint distribution of (Y_1, Y_2). We may decompose the joint distribution of (Y_S, S) into the conditional distribution of Y_S given S and the marginal distribution of S. The conditional distribution of Y_S given S is specified by the following probabilities:

$$t_{k\ell} = P(Y_\ell = k | S = \ell), \quad \ell = 1,2 \text{ and } k = 0,1 . \tag{7}$$

The marginal distribution of S is specified by the following probabilities

$$P(S = \ell), \ \ell = 1,2 . \tag{8}$$

The fundamental problem in a population-level causal inference (and therefore of all stronger forms of causal inferences) is the estimation of $q(\ell)$ for $\ell = 1,2$. However, the only data we can obtain in a causal study allows us to estimate the conditional probabilities given in (7). Thus, a question of paramount importance in causal inference is: when are $q_k(\ell) = P(Y_\ell = k)$ and $t_{k\ell} = P(Y_\ell = k | S = \ell)$ equal? That is, we are led to seek conditions under which the following equation holds:

$$P(Y_\ell = k | S = \ell) = P(Y_\ell = k) . \tag{9}$$

There are two very important cases where equation (9) holds — random assignment and homogeneous populations. We discuss each of these briefly in turn.

Random assignment: If S is statistically independent of Y_ℓ, then equation (9) must hold by definition of statistical independence. How can S be made to be independent of Y_ℓ? There is no way to be absolutely sure that S is independent of Y_ℓ. However, the process of "random" assignment of the values of S to the units in Q makes it plausible to assume that equation (9) holds if Q is large. Thus, under randomization we have

$$P(Y_1 = k, Y_2 = k' | S = \ell) = P(Y_1 = k, Y_2 = k') \qquad (10)$$

and equation (9) follows. The statistical independence expressed in (10) is a very important point in the justification of randomization but it is apparently not appreciated by numerous writers on the subject. For example, it is often asserted that there is a "difficulty" in resolving randomization and the Bayesian/likelihood/modelling framework (Basu, 1980; Kempthorne, 1976; Kruskal, 1980). However, equation (9) is a fundamental one for both Bayesians and frequentists because it makes a parameter that can be estimated from data (i.e. $P(Y_\ell = k | S = \ell)$) equal to the causal parameters of interest, $q_k(\ell)$.

One source of confusion is that equation (10) does not imply that the observed value Y_S is independent of S. That is, the following equation does not hold in general:

$$P(Y_S = k | S) = P(Y_S = k) \qquad (11)$$

Equation (11) does hold under the null hypothesis that $P(Y_\ell = k)$ does not depend on the level of exposure, ℓ. Of course, this fact is usually not of much interest to the Bayesian who wants to summarize evidence in the data about causal effects by the posterior distribution of causal parameters.

Homogeneous population: If Q is a homogeneous population, then equation (9) holds trivially without any assumption about the dependence or independence of S and Y_ℓ. This is because in a homogeneous population Y_ℓ is constant over units and constants are always independent of every random variable. Thus, for homogeneous populations randomization is not necessary for drawing population-level causal inferences.

In a nonrandomized study, it is often not believeable that S is statistically independent of Y_ℓ so that equation (9) may not hold. Thus, in a nonrandomized study the observed values of Y_ℓ are not representative of the marginal distribution of Y_ℓ over all of Q. However, if Q is a homogeneous population, then equation (9) must hold trivially. Covariates defining subpopulations play a crucial role in nonrandomized studies of causal effects. First, the subpopulations defined by them can be nearly homogeneous in which case equation (9) almost holds within each. Second, within each subpopulation it may be plausible to accept the assumption of conditional independence between Y_ℓ and S; at best, there may be no data to contradict this assumption. The next section addresses this issue in more detail.

The Role of Covariates

Suppose that Q can be partitioned into strata on the basis of a covariate X. We may then consider the possibility that equation (9) holds in each X-stratum even though equation (9) does not hold for all of Q. That is, we may ask whether or not

$$P(Y_\ell = k | S = \ell, X=x) = P(Y_\ell = k | X=x) \qquad (11)$$

for all values of ℓ, k and x? As mentioned earlier, there are two reasons why we may be willing to assume (11) even if we are not willing to assume (9). The first occurs when X is an ideal covariate and all the X-strata are themselves populations with homogeneous responses. Then we know that (11) holds automatically. The second occurs when we know or are willing to assume that S and Y are independent given X. We may be willing to make this assumption for one of two reasons. The first is that we actually randomly assigned the values of S within each stratum. The second is that we may be willing to make this assumption because there is nothing in the data that will contradict it. This is a subtle point and one that needs to be elaborated. If we assumed that S had been randomly assigned and was therefore independent of Y_ℓ then this assumption could be immediately contradicted by looking at the distribution of X given S. If S had been randomly assigned, then X and S would be independent so that

$$P(X = x | S = \ell) = P(X = x). \qquad (12)$$

If we examined the distribution of X given $S = \ell$ and saw that it did vary with the value of ℓ, then we would have evidence that S was <u>not</u> randomly assigned over all of Q and therefore that equation (9) does not hold. However, if we assumed that S was randomly assigned within each X-stratum we could not then use the observed distributions of X given $S = \ell$ to disprove this assumption.

Now suppose that equation (11) holds. We may use it to obtain a basic formula for the causal parameter, $q_k(\ell)$. We have

$$q_k(\ell) = P(Y_\ell=k) = \sum_x P(Y_\ell=k|X=x)\ P(X=x) \qquad (13)$$

so that

$$q_k(\ell) = \sum_x P(Y_\ell=k|S=\ell,\ X=x)\ P(X=x) \qquad (14)$$

Equation (13) is a basic fact of probabilities. Equation (14) relates two quantities that can be estimated, i.e., $P(Y_\ell=k|S=\ell,\ X=x)$ and $P(X=x)$ to the causal parameters. Thus, if equation (11) holds we can estimate the causal parameters and draw population-level causal inferences.

3. Causal Inference in Retrospective Case-Control Studies

The structure of a retrospective case-control study is considerably different from the general prospective study discussed in Section 2. In a case-control study a population of people is divided into those who have a particular symptom or disease of interest (i.e., the "cases") and those who do not have the symptom or disease (i.e., the "controls"). Samples of cases and controls are selected from this population and information about each selected person is obtained to ascertain: (a) the level of exposure to the particular causal agent of interest and (b) other medically relevant information which may be used to define subpopulations of units.

The response variable for a case-control study is the dichotomous variable that indicates whether or not the unit is a "case" or a "control", i.e.,

$$Y_S = \begin{cases} 1 \text{ if unit is a case} \\ 0 \text{ if unit is a control.} \end{cases}$$

Case-control studies are <u>retrospective</u> because they begin at the end-point of a prospective study (i.e., observations of the response variable for each unit in the study) and then look back in time to discover the level of causal agent to which each unit has been exposed (i.e., the value of the selection indicator S). In addition to this fundamental difference between case-control and prospective studies, there are two other differences that should be mentioned. First, since the investigator can only collect data on prior exposure to the causal agents of interest, it is impossible to employ randomization to assign units to levels of the causal agent. Thus case-control studies are never randomized. Prospective studies, on the other hand, may or may not employ randomization depending on the amount of control that is possible. Second, the populations studied in case-control studies usually consist of survivors only, because it is often impossible to obtain comparable data on individuals who are deceased. This limitation can have a serious effect on the interpretability of the results of a case-control study. We shall assume for the moment that the populations considered are not subject to mortality. We shall return to this point in the discussion of the example in section 4.

Although in principle it is almost always possible to formulate a prospective version of a case-control study, it is often much more expensive than the case-control study. There are several reasons for this: (a) prospective studies often require large sample sizes especially when the "cases" are rare (e.g., when $Y_S = 1$ represents a rare disease), (b) prospective studies often involve long time spans before relevant data become available. Hence it is likely that case-control studies will always be an attractive possibility for many types of scientific investigations, especially in the early stages of the research. It is therefore important to know their limitations, to design them as well as possible and to analyze the data collected in such studies correctly. Our goal in the present paper is to illuminate all of these points by applying the model for causal inference developed in section to case-control studies.

The standard two-way table and why it is misleading

In analysing data from a case-control study, it is customary to form and draw conclusions from the two-way table of counts illustrated in Table 4. We assume that this table was formed by randomly sampling m_{1+} "cases" and m_{0+} "controls" from the population.

Table 4 about here

In Table 4, $m_{k\ell}$ is the number of units in the study for which $Y_S=k$ and $S=\ell$. For example, m_{12} is the number of "cases" in the study that were observed at the 2nd level of exposure to the causal

agent. Before examining this table of sample data, let us consider the population table that underlies it. This population table gives the population proportion of people for which $S=\ell$ among all those for which $Y_S = k$, $k = 0,1$, $\ell = 1,2$. These population values are denoted by

$$r_{k\ell} = P(S = \ell | Y_S = k) \tag{15}$$

and arrayed as a population table in Table 5. The sample ratio

$$\hat{r}_{k\ell} = m_{k\ell}/m_{k+} \tag{16}$$

estimates $r_{k\ell}$. We shall call the $r_{k\ell}$ the retrospective probabilities of the study.

Table 5 about here

In this development we must emphasize the importance of representing the observed value of the response as Y_S. For example in (15) it would be incorrect to condition $Y_\ell = k$ since Y_ℓ is the response made if exposed to treatment level ℓ, whereas Y_S is the observed response. Because Y_S is being conditioned on in Table 5, it is sometimes said that in a case-control study exposure is the dependent variable and diagnosis (i.e., case or control) is the independent variable. This description confuses the scientific question of interest, and we will not describe the situation in these terms.

		Level-of-exposure S=1	S=2	Total
"cases"	$Y_S=1$	m_{11}	m_{12}	m_{1+}
"controls"	$Y_S=0$	m_{01}	m_{02}	m_{0+}
Total		m_{+1}	m_{+2}	m_{++}

Table 4: The standard 2-way sample table showing the distribution of cases and controls observed at each level of exposure to the causal agent.

	Levels-of-exposure $S = 1$	$S = 2$	Total
"cases" $Y_S=1$	$r_{11} = P(S=1 \mid Y_S=1)$	$r_{12} = P(S=2 \mid Y_S=1)$	1
"controls" $Y_S=0$	$r_{01} = P(S=1 \mid Y_S=0)$	$r_{C2} = P(S=2 \mid Y_S=0)$	1

Table 5. The population table of retrospective probabilities r_{kl} that underlies the sample table in Table 4.

If we consider the weakest level of causal inference, i.e., a population-level of causal inference, then the causal parameters are the marginal probabilities $P(Y_1=1)$ and $P(Y_2=1)$. Thus, the retrospective probabilities in (15) are not, in themselves, of any causal interest, because, at the very least, they describe the wrong events. However, by applying the usual rules of probability, we may reverse the roles of S and Y_S in (15) and obtain more interesting probabilities. This reversal is the usual justification for ever looking at Table 4.

Relating retrospective and prospective probabilities

To reverse the roles of S and Y_S we make use of Bayes theorem to obtain

$$P(Y_S=k|S=\ell) = P(S=\ell|Y_S=k)\frac{P(Y_S=k)}{P(S=\ell)} \quad (17)$$

However,

$$P(Y_S=k|S=\ell) = P(Y_\ell=k|S=\ell) , \quad (18)$$

so it follows that,

$$P(Y_\ell=k|S=\ell) = P(S=\ell|Y_S=k)\frac{P(Y_S=k)}{P(S=\ell)} \quad (19)$$

or

$$t_{k\ell} = r_{k\ell}a_k b_\ell$$

where

$$a_k = P(Y_S=k)$$

$$b_\ell = 1/P(S=\ell) ,$$

and $t_{k\ell}$ is given in (7).

Hence, in order to transform the retrospective probabilities $r_{k\ell}$ in (15) and Table 5 into the more interesting "prospective" probabilities, $t_{k\ell}$, we need only multiply the entries in Table 5 by a row factor (i.e., a_k) and a column factor (i.e., b_ℓ). We have illustrated the array of "prospective" probabilities of (18) in Table 6.

Table 6 about here

Note that

$$P(S=\ell) = \sum_k P(S=\ell|Y_S=k)\ P(Y_S=k) = \sum_k r_{k\ell}\ \ a_k$$

so that the prospective probabilities in (19) can be calculated from knowledge of a) the retrospective probabilities $r_{k\ell}$, and b) the overall proportions of cases and controls in the population $a_k = P(Y_S=k)$.

The cross-product ratio for Table 5 may be expressed as:

$$\alpha^* = \frac{r_{12}}{r_{11}} \Big/ \frac{r_{02}}{r_{01}} = \frac{P(S=2|Y_S=1)}{P(S=1|Y_S=1)} \Big/ \frac{P(S=2|Y_S=0)}{P(S=1|Y_S=0)}\ ; \qquad (20)$$

and the cross-product ratio for Table 6 may be expressed as:

$$\alpha^{**} = \frac{P(Y_2=1|S=2)}{P(Y_2=0|S=2)} \Big/ \frac{P(Y_1=1|S=1)}{P(Y_1=0|S=1)} \qquad (21)$$

Because Tables 5 and 6 are related via row and column multiplication, it is well-known (e.g. Bishop, Fienberg, Holland (1975)) that

$$\alpha^* = \alpha^{**} \qquad (22)$$

		Level-of-exposure	
		S=1	S=2
"Cases"	Y=1	$P(Y_1=1 \mid S=1)$	$P(Y_2=1 \mid S=2)$
"Controls"	Y=0	$P(Y_1=0 \mid S=1)$	$P(Y_2=0 \mid S=2)$
Total		1	1

Table 6. The population table of prospective probabilities that may be derived from Table 5 by row and column multiplication.

Population-level causal inferences

Now let us return to the question of making a population-level causal inference about the effect of the causal agent on the probability of becoming a "case." The parameters of interest in such a causal inference are the causal parameters $q_k(\ell) = P(Y_\ell=k)$ or, equivalently, the odds associated with these probabilities, i.e.

$$\beta(\ell) = \frac{P(Y_\ell=1)}{P(Y_\ell=0)} , \qquad (23)$$

$\ell = 1,2$. The odds in (23) for $\ell=2$, relative to $\ell=1$ gives the odds ratio

$$\alpha = \frac{\beta(2)}{\beta(1)} = \frac{P(Y_2=1)}{P(Y_2=0)} \Big/ \frac{P(Y_1=1)}{P(Y_1=0)} = \frac{q_1(2)}{q_0(2)} \Big/ \frac{q_1(1)}{q_0(1)} \qquad (24)$$

Even though α represents less information than both $\beta(1)$ and $\beta(2)$, interest often focuses on the odds ratio in case-control studies. Certainly α does give a measure of the change in $q(2)$ relative to $q(1)$.

If we could assume that S and (Y_1,Y_2) were independent, then it would follow from (21) that α and α^* would be equal. This would justify examining Table 4, because the cross-product ratios directly estimated by this table (i.e., α^*) would be equal to the cross-product ratio of the causal parameters (i.e. α). However, case-control studies are non-randomized studies so that randomization can not be a generally satisfactory basis for assuming that S and (Y_1,Y_2) are independent. Furthermore, by examining the distribution of a covariate X given $S=\ell$, we can often convince ourselves in a case-control study that S was not even approximately randomized. Thus it is essential in case-control

studies to examine more detailed aspects of the data than those which are summarized by Table 4 in order to have some hope of drawing reasonable conclusions.

The odds ratio α^* in a case-control study may not equal α due to the self-selection of individuals into exposure categories. We conclude that basing the analysis of a case-control study on Table 4 is potentially misleading because it ignores the possibility of bias due to self-selection into the exposure conditions. We hasten to point out that since population-level causal inferences are the weakest of the three types of causal inferences we discussed in section 2, it follows that if population-level causal inferences are impossible from the data in Table 4, so are all other types of causal inferences.

The role of covariates in case-control studies

If there is a covariate (or set of covariates) X which is measured on each unit in the study, then we may form a table like Table 4 for each value of X. Let $m_{k\ell x}$ be the number of units in the study for which $Y_S = k$, $S = \ell$ and $X = x$. These are arrayed in Table 7 for X=x.

Table 7 about here

The ratios

$$\hat{r}_{k\ell x} = m_{k\ell x}/m_{k\ell x} \tag{25}$$

estimate the population retrospective probabilities

$$r_{k\ell x} = P(S=\ell | Y_S=k, X=x). \tag{26}$$

Value of X=x

	Level-of-exposure S = 1	S = 2	total
"cases" Y_S=1	m_{11x}	m_{12x}	m_{1+x}
"controls" Y_S=0	m_{01x}	m_{02x}	m_{0+x}
total	m_{+1x}	m_{+2x}	m_{++x}

Table 7: The distribution of cases and controls in the sample observed at each level of exposure to the causal agent, for X = x.

We may apply Bayes theorem to reverse the roles of S and Y_S in (26) as we did in (17). This yields the following equation

$$P(Y_\ell=k \mid S=\ell,\ X=x) = r_{k\ell x}\, a_{kx}\, b_{\ell x} \tag{27}$$

where

$$a_{kx} = P(Y_S=k \mid X=x) \tag{28}$$

and

$$b_{\ell x} = 1/P(S=\ell \mid X=x)\ : \tag{29}$$

By the same argument given for α^* and α^{**} we have

$$\alpha_x^* = \alpha_x^{**} \tag{30}$$

where

$$\alpha_x^* = \frac{r_{12x}}{r_{11x}} \Big/ \frac{r_{02x}}{r_{01x}} \tag{31}$$

and

$$\alpha_x^{**} = \frac{P(Y_2=1 \mid S=2,\ X=x)}{P(Y_2=0 \mid S=2,\ X=x)} \Big/ \frac{P(Y_1=1 \mid S=1,\ X=x)}{P(Y_1=0 \mid S=1,\ X=x)} \tag{32}$$

If it is reasonable to assume that (Y_1,Y_2) and S are conditionally independent given X=x, then

$$\alpha_x^* = \alpha_x \qquad (33)$$

where

$$\alpha_x = \frac{P(Y_2=1|X=x)}{P(Y_2=0|X=x)} \Big/ \frac{P(Y_1=1|X=x)}{P(Y_1=0|X=x)} \qquad (34)$$

On the other hand, the cross-product ratio that is determined by the causal parameters is α in equation (24). The relationship between α and the values of α_x is not a simple one due to the nonlinear form of the cross-product ratio. For example, the average value of α_x over the distribution of X does not equal α in general. There is no simple analogue to formula (14) for the retrospective cross-product ratios.

Suppose $\alpha_x^* = \alpha_0^*$ for all values of x. If this happens then we shall say that the data in the case-control study exhibit a constant cross-product ratio -- i.e., α_x^* is constant across all values of x. If we are willing to further assume that (Y_1,Y_2) and S are conditionally independent given X then

$$\alpha_x = \alpha_x^* = \alpha_0 \qquad (35)$$

Even when (35) holds there is still no simple relationship between α_0 and α. The general formula relating α_0 to α is given in (36).

$$\alpha = \alpha_0 \left[\frac{\sum_x \left(\frac{q_1(1|x)}{q_0(1|x) + \alpha_0 \, q_1(1|x)} \right) P(X=x)}{\sum_x \left(\frac{q_0(1|x)}{q_0(1|x) + \alpha_0 \, q_1(1|x)} \right) P(X=x)} \Big/ \frac{q_1(1)}{q_0(1)} \right] \qquad (36)$$

where

$$q_k(\ell|x) = P(Y_\ell = k|X=x) \; . \qquad (37)$$

We note that the causal parameters $q_k(\ell)$ appear in (36) along with their conditional versions $q_k(1|x)$. The example in Table 8 shows that α_0 and α need not be equal.

Table 8 about here

All is not lost however, because α_0 is a causally interesting quantity itself. It is the amount by which the odds for $Y_2 = 1$ is increased over the odds for $Y_1 = 1$ in each X-stratum of Q. Thus α_0 is a useful parameter to estimate because it has causal relevance in each of the subpopulations of Q. Since α_0 is specific to the X-strata of Q, it provides causal inferences about the effects of the levels of the causal agent in Q that are at a more detailed level than population-level causal inferences. However it is not as strong as the intermediate-level, or the unit-level causal inferences discussed earlier in section 2.

Table 8: Example showing that α_o^* and α need not be equal

X takes on two values X=1, X=2

$\ell = 1,2;\ k = 0,1$

X=1

	$P(Y_1=k\|X=1)$	$P(Y_2=k\|X=1)$
k=1	1/10	5/10
k=0	9/10	5/10
Total	1	1

$\alpha_1^* = 9$

X=2

	$P(Y_1=k\|X=2)$	$P(Y_2=k\|X=2)$
k=1	1/100	1/12
k=0	99/100	11/12
Total	1	1

$\alpha_2^* = 9$

therefore $\alpha_o^* = 9$

If P(X=1) = .1 and P(X=2) = .9 then

	$P(Y_1=k)$	$P(Y_2=k)$
k = 1	19/1000	150/1200
k = 0	981/1000	1050/1200
Total	1	1

and $\alpha = 7.4$

If P(X=1) = .5 and P(X=2) = .5 then

	$P(Y_1=k)$	$P(Y_2=k)$
k=1	11/200	70/240
k=0	189/200	170/240
Total	1	1

and $\alpha = 7.1$

Our conclusion is that in a case-control study the simple 2-way table (Table 4) usually holds no causal interest. The only hope is to stratify on covariates and to estimate the α_x^*. If the stratified table exhibits constant cross-product ratios then the strongest form of causal inference appears to be to estimate α_0^* and assume that it equals α_0. These latter parameters give the amount that the second level of the causal agent increases the proportion of units in each X-stratum that are "cases" relative to the first level of the causal agent. This "amount of increase" is in terms of the odds corresponding to the proportions. Thus, for example, for a given value of the proportion $P(Y_1=1|X=x)$, we calculate $P(Y_2=1|X=x)$ via the formula

$$P(Y_2=1|X=x) = \frac{\alpha_0 \, P(Y_1=1|X=x)}{(P(Y_1=0|X=x) + \alpha_0 \, P(Y_1=1|X=x))} \quad (38)$$

Comparing this to the given value of $P(Y_1=1|X=x)$ leads to a causal inference about the effect of the causal agent when X=x.

Prospective vs. retrospective matching

Another way to see the fundamental weakness in retrospective studies is to compare prospective matching, which matches an exposed and unexposed unit with respect to X, and retrospective matching which matches a case and a control with respect to X. Suppose that S and Y_1, Y_2 are independent given X, so that at each level of X we have a randomized experiment, i.e. the experiment is a randomized block with blocks defined by X. Prospective matching reconstructs the randomized block experiment by creating matched pairs of exposed and unexposed units. The average matched pair difference is an unbiased estimate of the treatment effect for the population defined by the values of X in the matched pairs. Thus

prospective matching on X perfectly controls for X whenever both members of each matched pair have the same values of X.

In contrast, retrospective matching on X in general cannot perfectly control for X because it does not reconstruct the randomized block experiment. In each matched pair, one member is a case and one member is a control; to reconstruct the randomized block experiment, one member must be exposed and one unexposed, which generally does not occur when one member is a case and the other a control. Thus summaries from the case-control matched sample such as the cross-product ratio do not represent an estimate for which X has been controlled, even when all matched pairs are exactly matched with respect to X. With retrospective matches, we really need to estimate the cross-product ratio in each matched pair, and this requires building a model relating Y_1, Y_2 to X and S. We illustrate this in the next section.

4. An Example

The following data are taken from a case-control study of the relationship of coffee drinking and occurances of myocardial infarctions (MI) by Jick et al (1973). We use these data for illustrative purposes only. A total of 24,741 patients were classified as "cases" (had an MI) or "controls" (did not have an MI). Table 9 shows the standard 2-way table that presents the cases and controls cross-classified by the potential causal agent under study -- self reported daily coffee consumption. Although our previous notation has considered only two levels of the causal agent, Table 9 presents four levels; the extensions needed to handle this extra complexity are simple. The cross-product ratios estimated in Table 9 are defined by

$$\alpha^*(\ell) = \alpha^{**}(\ell) = \frac{P(Y_\ell = 1|S = \ell)}{P(Y_\ell = 0|S = \ell)} \Big/ \frac{P(Y_1 = 1|S = 1)}{P(Y_1 = 0|S = 1)} \qquad (38)$$

Table 9 about here

Table 9 suggests a modest increase in the risk of MI among persons who drink coffee. The odds ratios range from 1.5 to 1.8. The cross-product ratios exhibited in Table 9 are not monotone in the amount of self-reported coffee drinking and the effect seems to be almost as strong for persons who drink 1-2 cups per day as for those who drink 6+ cups per day.

However, Table 9 does not take various background variables into account and, as we have discussed earlier, therefore is likely to be misleading because (1) it is not reasonable to believe that the drinking of coffee is independent of other relevant factors, and (2) the matching of cases and controls on background variables does not control for them. In addition to the variable S = level of self-reported coffee intake and Y = case or control, the following set of variables were also available on all patients in the study.

A = Age: 6-levels: 20-29, 30-39, ..., 70-79.

G = Gender: 2-levels: Male, Female.

C = Smoking: 3-levels: other, ex-smoker, current smoker.

O = Other heart disease: 2-levels: yes, no.

In addition the data were collected from 24 suburban Boston hospitals so that a fifth variable, H = hospital, was included in the analysis (with 24 levels). This resulted in a covariate X which takes on 6x2x3x2x24 = 1728 values so that when table 9 is stratified on X we obtain a 7-way contingency table with 2x4x1728 = 13824 cells. With a total of 24,741 observations, this gives us about 1.9 observations per cell -- a very

Table 9: Cross-Tabulation of self-reported coffee intake (S) by cases and controls (Y) for 24,741 patients

S = Self-reported coffee consumption per day

		S=1 0 cup/day	S=2 1-2 cups	S=3 3-5 cups	S=4 6+ cups	Total
$Y_S = 1$	MI cases	128	269	147	86	630
$Y_S = 0$	non-MI controls	6918	9371	5290	2532	24111
	Total	7046	9640	5437	2618	24741

Estimated raw cross-product ratios $\alpha^*(\ell)$ relative to $\ell=1$

$\hat{\alpha}^*(2)$	$\hat{\alpha}^*(3)$	$\hat{\alpha}^*(4)$
1.551	1.502	1.836

sparse table indeed! Many approaches to simplifying this sort of situation are possible. We shall use log-linear contingency table models (a) because of their direct relationship to the cross-product ratios, (b) because they allow us to see the effect of all of the covariates simultaneously and (c) because they do not force us to rely heavily on the sparse 7-dimensional table.

Loglinear Models for this Problem

Let X = (A,G,C,O,H) denote our complete vector of covariates. The retrospective probabilities $r_{k\ell x}$ from (25) may be expressed as:

$$\log (r_{k\ell x}) = u + u_{1(k)} + u_{2(\ell)} + u_{3(x)} + u_{12(k,\ell)} + u_{13(k,x)} + u_{23(\ell,x)} + u_{123(k,\ell,x)} \quad (39)$$

where the u-terms in (39) are assumed to satisfy the usual ANOVA-like identifying constraints $u_{1(+)} = u_{2(+)} = 0$ etc. We need to express the cross-product ratios

$$\alpha_x^*(\ell) = \frac{r_{1\ell x}}{r_{11x}} \Big/ \frac{r_{o\ell x}}{r_{olx}} \quad (40)$$

in terms of the u-terms in (39). It is easy to show that the following equation holds,

$$\alpha_x^*(\ell) = \alpha^*(\ell) \exp\{u_{123(1,\ell,x)} - u_{123(1,1,x)} - u_{123(o,\ell,x)} + u_{123(o,1,x)}\} \quad (41)$$

where

$$\alpha_o^*(\ell) = \exp\{u_{12(1,\ell)} - u_{12(1,1)} - u_{12(o,\ell)} + u_{12(o,1)}\}. \quad (42)$$

From (41) it follows that the hierarchical log-linear model (see, for example, Bishop, Fienberg and Holland, 1975) specified by setting $u_{123} = 0$ corresponds to the assumption that

$$\alpha_x^*(\ell) = \alpha_o^*(\ell) \quad (43)$$

for all x. Thus, we may investigate the question of whether or not the cross-product ratios $\alpha^*_x(\ell)$ depend on x by testing three-way interactions of the various covariates in X and with Y_S and S. Furthermore, if a model where $u_{123} = 0$ is acceptable, the estimated u_{12}-terms may be used to obtain estimates of $\alpha^*_o(\ell)$. If we are willing to make the assumptions necessary to insure that $\alpha^*_o(\ell) = \alpha_o(\ell)$, where $\alpha_o(\ell)$ is the causally relevant parameter discussed in Section 3, then we may test $\alpha_o(\ell) = 1$ (i.e., no effect of different levels of the causal agent) by testing that $u_{12} = 0$. This test will adjust for the distribution of the covariates in the several exposure groups.

Simplifying the analysis

As described above it may seem as though we are considering the whole 2x4x1728 table, but one important feature of the use of log-linear models is that they do not force this unless there is sufficient data to do so. Instead we break up X = (A,G,C,O,H) into various marginal distributions and expand the model in (39) to make use of them. In the present example we expanded the table to the full seven-dimensions, but only fit effects for the following pairs and triples of variables:

(u_{12}) SY/

(u_{23}) HS/AS/GCS/GOS/COS/

(u_{13}) HY/AGY/ACY/AOY/GCY/GOY/COY/

(u_3) HA/HG/HC/HO/AGC/AGO/ACO/GCO/.

The u-terms in parenthesis indicate which terms in (39) have been expanded in the 7-way table.

Results

If we fit the log-linear model indicated by the pairs and triples of variables in (44) and then delete the SY terms and refit the model, we obtain a likelihood ratio test of $\alpha_o(\ell) = 1$. The value of the likelihood ratio statistics is 12.3 which under the null hypothesis has 3 degrees of freedom. Thus, this analysis results in a significant relationship between coffee-consumption and myocardial infarctions. The estimated $\alpha_o(\ell)$ values are

$\hat{\alpha}_o(2)$	$\hat{\alpha}_o(3)$	$\hat{\alpha}_o(4)$	
1.188	1.235	1.719	(45)

as opposed to the raw cross-product ratios given in Table 9. These adjusted cross-product ratios are monotonic in the amount of coffee consumed and the major effect is seen to be for high levels of coffee consumption.

To study the question of whether $\alpha_x(\ell)$ varies with x we fit 5 additional models each of which replaces SY in (44) by one of these triples of variables: HSY, ASY, GSY, CSY, or OSY. The likelihood ratio statistics for these models, the degree of freedom and attained significance levels are given in Table 10.

Table 10 about here

None of these interactions are strong enough to be statistically significant. This result contradicts previous analysis of these data that found an interaction with these variables, (Miettinen, O.S., 1976).

Table 10

Summary of study of dependence of $\alpha_x(\ell)$ on x

Interaction with	df	ΔG^2	Level attained
H	69	79.39	.20
A	15	10.31	.80
G	3	2.5	.47
C	6	8.83	.18
O	3	3.97	.25

REFERENCES

Basu, D., "Randomization Analysis of Experimental Data: The Fisher Randomization Test", Journal of the American Statistical Association, 1980.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W., Discrete Multivariate Analysis: Theory and Practice, The MIT Press, Cambridge, Mass., 1975.

Cochran, W.G., "The Planning of Observational Studies of Human Population," Journal of the Royal Statistical Society, Series A, 128, Part 2, 234-255, discussion 255-265, 1965.

Cornfield, J. "A method of estimating comparative rates from clinical data, application to cancer of the lung, breast and cervix", Journal of the National Cancer Institute, 11, 1269-1275, 1951.

Cornfield, J. "A statistical problem arising from retrospective studies" Proceedings of the Third Berkeley Symposium, 4, 135-148, 1956.

Cox, D.R., Planning of Experiments, John Wiley & Sons, Inc., New York, 1958.

Fisher, R.A., The Design of Experiments, Oliver & Boyd, Edinburgh, 1935.

Granger, C.W.J., "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," Econometrica, 37, 424-438, 1969.

Jick, H. et. al. "Coffee and Mycardial Infarcation", New England Journal of Medicine, 289, 63-67, 1973.

Kempthorne, O., The Design and Analysis of Experiments, John Wiley & Sons, Inc., New York, 1952.

Kempthorne, O., Discussion of "On rereading R.A. Fisher by Leonard J. Savage," The Annals of Statistics, 4, 495-497, 1976.

Kruskal, W., "The Significance of Fisher," Journal of the American Statistical Association, 1980.

Mieltinen, O.S. "Stratification by a multivariate confounder score", American Journal of Epidemiology, 104, 609-620, 1976.

Rubin, D.B., "Bayesian Inference for Causal Effects: The Role of Randomization," The Annals of Statistics, 6, 1, 34-58, 1978.

Rubin, D.B., Discussion of "Randomization Analysis of Experimental Data: The Fisher Randomization Test," Journal of the American Statistical Association, 1980.